We thank the reviewers for insightful feedback and suggestions. We will clarify our figures and text accordingly.

Intended role of theory and experiments (R4). The goal of our theory is to show that non-leaking augmentations *do not inherently harm* the training objective — the results would inevitably degrade if the equilibrium point was affected by the augmentations. However, we agree that the equilibrium is hardly ever achieved in practice, and that the effectiveness of augmentations ultimately depends on the complex interaction between many aspects of the training process. Thus, in order to *demonstrate the benefits* of our technique, we rely on extensive practical experiments and place considerable emphasis on comparing against a sufficiently large set of alternative approaches, including e.g. adaptive dropout (Figure 8a).

The reason why augmentations help in the adversarial game is that they make it harder for the raw D outputs of real and generated images to drift apart, as visualized in Figures 1 and 6. This is important because the gradients that G receives from D become meaningless once the overlap between the distributions is lost.

Early stopping (R4). As can be seen in Figure 6a, our method typically leads to monotonic convergence that clearly surpasses the best FID achievable using early stopping (Figure 1a). In the context of GANs, it is customary to report the lowest FID seen over the course of training and we also follow the same protocol. In this sense, our experiments already employ early stopping — largely to the benefit of the comparison methods. With ADA, this would not be strictly necessary since we could get comparable results by only looking at FID toward the end of the training in most cases.

Discriminator capacity (R1, R4). We agree with the reviewers that a sweep over D capacity would provide valuable insight and will gladly include it in the final version. That said, we have not observed that decreasing the capacity would prevent overfitting with small training sets ($\sim$2k), but it might reduce or postpone it slightly with moderately-sized ones ($\sim$30k). Similarly, we have not observed significant benefits from increasing the capacity, either.

Spectral normalization (R3). The StyleGAN2 paper also reports that spectral normalization did not help, in line with our results in Figure 8a. We do not know exactly why this is the case but would like to emphasize that the interaction between various regularization techniques and architectural choices is not fully understood yet. We suspect that the effectiveness of spectral normalization is tied to a specific kind of training setup that is sufficiently different from the one used in StyleGAN2. For example, papers where spectral normalization is shown to be beneficial do not typically employ explicit gradient penalty terms, such as R_1.

Different loss functions (R3). The exact behavior of D certainly depends on the loss function. We chose to focus on the original non-saturating logistic loss that was used in StyleGAN2. We suspect that hinge loss would exhibit comparable behavior, as the shape of the two functions is substantially similar: $f(x) = \log(\mathrm{sigmoid}(x))$ vs. $f(x) = -max(0, 1-x)$. WGAN and WGAN-GP are somewhat trickier, because the outputs of D are not "grounded" to any particular range, so their mean and standard deviation tend to drift around over the course of training. Nevertheless, we would generally expect D_{train} to stay above $D_{\text{generated}}$, with the difference becoming more pronounced when D starts to overfit.

D-only augmentation (R1). Let us consider what would happen if the augmentations were only applied when training D, but skipped when training G. In this case, D would see the true distribution of generated images ($\mathbf{x}$) when G is being trained and the augmented distribution ($\mathcal{T}\mathbf{x}$) when D itself is being trained. We have tested this variant and observed that the mismatch between these two distributions leads to an immediate mode collapse. In effect, D is only trained to guide $\mathcal{T}\mathbf{x}$ toward $\mathcal{T}\mathbf{y}$, so it is unable to provide meaningful gradients for guiding $\mathbf{x}$ toward $\mathbf{y}$. The situation is markedly different with bCR, because the main training objective of D is still based on the true distributions — the augmented distributions are used only in the auxiliary loss terms, so their effect is weaker and less direct.

Fluctuations in Figure 5d (R1). We have noticed that non-adaptive discriminator augmentation tends to cause fluctuation in the training dynamics once D has entered the overfitting regime, i.e., when FID has started increasing. This often happens toward the end of the training when the specific choice of fixed p is no longer sufficient to prevent overfitting. We suspect that the fluctuation is caused by D having become overly sensitive to a small set of image features and reacting strongly to the stochastic effect of the augmentations on these features.

Clarifications (R1). In Figure 2, the yellow boxes indicate the loss function and the green boxes indicate the network being trained. In the ADA heuristic, we adjust p every 4 minibatches simply because of the way the StyleGAN2 training loop is laid out; the results are not sensitive to this particular choice. Regarding the AFHQ dataset, we have tested the CAT and WILD categories and observed high-quality results comparable to the DOG category. We originally left these results out to save space but could include them in the final version.